LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 8 (2014), pp. 192-208
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24602

Review of Gabmap: Doing Dialect Analysis on the Web

Reviewed by Conor Snoek, *University of Alberta*

1. OVERVIEW.[1] Dialectometry is the application of quantitative methods - numerical taxonomy - to the study of geo-linguistic variation (Goebl 2005: 498). It encompasses a range of methods for measuring and visually representing linguistic data, especially as regards the geographic distribution of linguistic variables. The discipline emerged as a specialized field in the 1970s with the work of French linguist Jean Séguy. The methodology developed in the field has broader implications for the study of the grouping of languages and language varieties, however. The application that I will review here, Gabmap[2], is a package uniting a set of computational methods and cartographic tools that can be useful to researchers more broadly interested in the distribution of linguistic variables, and the classifications that can be based on them. Recent years have seen a sharp increase in the use of quantitative methods in linguistics, especially through their implementation in the statistical programming environment R[3]. Indeed, many of the functions that Gabmap makes available, are also accessible in R (some of Gabmap's functions are even are implemented in R). Nevertheless, R remains unattractive to many researchers because of its interface and steep initial learning curve. Gabmap provides easy access to a selection of computational methods and provides guidance for their correct usage. Furthermore, it unites techniques used for the distance-based grouping of linguistic data with the possibility of representing those data on geographic maps. This combination of features makes it an attractive tool that stands to increase interest in quantitative methods and generate interesting research.

Gabmap is a free, web-based application that offers access to a number of tools for measuring distance between languages or dialects, plotting those distances on maps, and drawing cluster dendrograms. It is specifically designed so that even researchers with limited experience in the use of computer-aided data analysis software can make use of the program (Nerbonne et al. 2011). Gabmap was designed with the idea of making dialectometry, the quantitative development of dialectology, more accessible. Ideally, this software would further research on variation and the geographic distribution of linguistic features. While Gabmap is essentially designed with the evaluation of phonetic strings in mind, the application can be used in the broader sense of locating distributions of linguistic variables on geographic maps.

Having first heard about Gabmap from a visiting doctoral student at the University of Alberta (Martijn Wieling), I had the opportunity of participating in a workshop on using Gabmap in Gothenburg in 2011. The workshop was brief, but very useful. The most

[1] Gabmap was developed at the Department of Information Science of the University of Groningen by Peter Kleiweg (original implementation), Therese Leinonen (user documentation), Charlotte Gooskens, Rinke Colen, Çağrı Çöltekin and John Nerbonne

[2] Gabmap can be accessed at http://www.gabmap.nl/.

[3] The R Project for Statistical Computing http://www.r-project.org/.

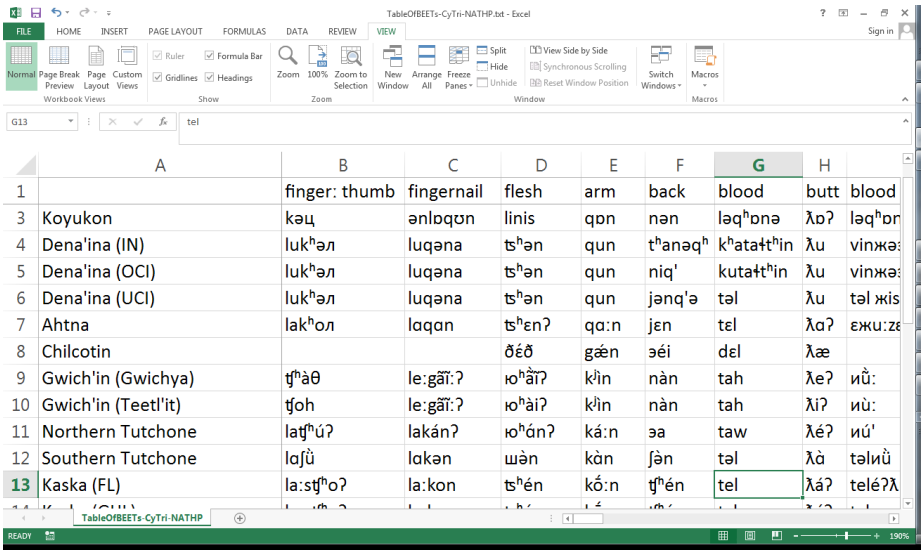
E-ISSN 1934-5275

tedious aspect of using Gabmap is the preparation of the dataset, which I will explain in more detail below. In this review I will be showing examples of my own work comparing phoneme strings denoting body-part terms in Northern Athapaskan languages. The application of dialectometrical tools is appropriate for these Athapaskan languages because their relative status is poorly established. In fact, Northern Athapaskan has been viewed as a dialect complex whose patterns of sound correspondences do not resolve into a tree-like structure (Krauss and Golla 1981). While it is not the case that all the Athapaskan languages of this area are dialects in the sense of being mutually intelligible, the distinctions are uncertain enough to warrant their exploration through measures of aggregate phonological distance (this notion will be explained in §5). The example data are chosen to demonstrate the functions of Gabmap with data not available in the demos on the website and to show that Gabmap can be usefully applied to problems not strictly within the traditional domain of dialectology.

The developers of Gabmap have provided a demo in which a range of functions is introduced to first-time users. Following the link on the start-up page will bring the user to two pre-compiled datasets. At the time of this writing these datasets were from a study of Pennsylvania English and Dutch dialects in The Netherlands and Flanders. Once a demo project has been selected, the user is free to investigate by choosing one of the options for data visualization and analysis. However, the user is well-advised to go through the tutorial first, or read the paper by Nerbonne et al (2011) available via the *Publications* link at the bottom of the start-up page.

The start-up page also has a number of buttons in a horizontal navigation bar just below the header: *News*, *Docs*, *Repository*, *Events* and *About*. However, this part of the page is not well developed. While there is some useful information here, the most promising link, entitled *Manual*, appears to contain roughly the same information as the *Tutorial*. The latter, however, is very useful and highly recommended for first time users.

2. PREPARING THE DATA. Gabmap is designed to handle four types of data: string data, numeric data, categorical data and difference data. The data needs to be entered in the format of a text file (.txt) that has the conceptual categories for the items across the horizontal axis and the languages (or varieties) to be compared along the vertical axis. Figure 1, below, is an example of a dataset in which a number of Athapaskan languages are compared on the basis of a list of body-part terms. Once the data has been arranged so that the terms under comparison are listed horizontally across the top row, and the languages in a vertical column at the far left, the data can be exported to a text file (.txt). Microsoft Excel offers the option to export under Save as > Unicode text (.txt), resulting in the tab delimited format displayed in Figure 1.



	A	B	C	D	E	F	G	H	
1		finger: thumb	fingernail	flesh	arm	back	blood	butt	blood
3	Koyukon	kəɻ	ənlɒqʊn	linis	qɒn	nən	ləqʰɒnə	ƛɒʔ	ləqʰɒn
4	Dena'ina (IN)	lukʰəл	luqəna	ʦʰən	qun	tʰanəqʰ	kʰataɬtʰin	ƛu	vinжəɜ
5	Dena'ina (OCI)	lukʰəл	luqəna	ʦʰən	qun	niq'	kutaɬtʰin	ƛu	vinжəɜ
6	Dena'ina (UCI)	lukʰəл	luqəna	ʦʰən	qun	jənq'ə	təl	ƛu	təl жis
7	Ahtna	lakʰол	lɑqɑn	ʦʰɛnʔ	qɑ:n	jɛn	tɛl	ƛɑʔ	ɛжu:zɛ
8	Chilcotin			ðéð	gǽn	ɘéi	dɛl	ƛæ	
9	Gwich'in (Gwichya)	ʧʰàθ	le:gãĩ:ʔ	ʁoʰãĩʔ	kʲìn	nàn	tah	ƛeʔ	иǔ:
10	Gwich'in (Teetl'it)	ʧoh	le:gãĩ:ʔ	ʁoʰàiʔ	kʲìn	nàn	tah	ƛiʔ	иù:
11	Northern Tutchone	laʧʰúʔ	lakánʔ	ʁoʰánʔ	ká:n	ǝa	taw	ƛéʔ	иú'
12	Southern Tutchone	lɑʃǔ	lɑkǝn	ɯǝn	kǝn	ʃǝn	tǝl	ƛà	tǝlиǔ
13	Kaska (FL)	la:stʰoʔ	la:kon	ʦʰén	kǒ:n	ʧʰén	tel	ƛáʔ	teléʔƛ

FIGURE 1. Data in tab-delimited text format (Unicode)

Something to watch out for here is that Gabmap has pre-determined ways of dealing with data. It is therefore highly recommended to make use of the data inspection options (explained in more detail in §3). The dataset used here, for instance, has frequent occurrences of digraphs and trigraphs, and it is worth checking how these are handled by Gabmap. Athapaskan has several ejective series which are rendered in IPA with /t'/ or /tθ'/, for example. Gabmap treats each IPA symbol as its own character however, so that distances may be exacerbated unnecessarily if the user intends affricates to be considered as unitary characters. Furthermore, the apostrophe used to signify ejectivized consonants is also treated as an individual symbol resulting in undesired distortions. Because of that, it can be useful to replace all digraphs and trigraphs with unitary symbols in cases where this is phonologically justified (any Unicode character will do here, since the algorithm is taking only identity or difference into consideration).

In order to map the data geographically, the user needs to supply a map with a set of locations: one for each of the language varieties to be compared. Drawing the map can involve a certain amount of effort. Gabmap is designed to handle geographic data encoded in Keyhole Markup Language (KML), which is "an XML language focused on geographic visualization, including annotation of maps and images" (Open Geospatial Consortium 2014). KML files are most easily created using Google Earth[4]. The user interface in Google Earth provides functions with which the user can add annotations and overlays to geographic areas. These annotations are saved along with the geospatial data from Google Earth in file format called .kmz.

Creating the .kmz files necessitates gaining some familiarity with Google Earth. The time spent learning about some of the basic functions available in Google Earth is well invested, however, since it ultimately allows for the data to be plotted on maps tailored to

[4] The software is freely available at http://www.google.com/earth/.

the user's needs. Such user-created maps are especially useful if the data to be plotted do not correspond to units that represent modern political entities. For the dataset used here, the map encompassed parts of western Canada and Alaska.[5]

Google also offers tutorials on various functions on its webpages. For the purposes of creating a .kmz file for Gabmap, knowledge of just two Google Earth functions is entirely sufficient: *Add Placemark* (indicated by the yellow thumb tack icon) and the immediately adjacent *Add Polygon*. The former function allows for specific locations to be added, e.g. the community where the language under study is spoken or other relevant geographical point that is to form an item for comparison. The *Add Polygon* feature is used to draw a map outline by marking a series of points. Once the polygon has been drawn and the locations added, the information can be exported (*Save Places As…* for Windows users) and saved in an opportune location. In this process it is crucial that the names of the locations match exactly the names of the languages in the dataset, since this is the key that Gabmap uses to associate linguistic and geographic data. If there is any incongruity between the place names and the language names Gabmap will complain and no analysis will be carried out. Fortunately, correcting files and uploading them again is a painless process, thanks to the uncomplicated user interface.

Once the data has been organized and the map drawn and saved, the user needs to create a new account on the Gabmap website. The server that hosts Gabmap also stores user data. The user account could potentially remain open for long periods of time. Users will periodically be reminded of its existence. Failure to respond to the emails results in the deletion of the account. After registering the account the user is invited to upload the relevant files. As a last step before the data are analyzed, Gabmap must be informed about the nature of the variables: string (IPA transcriptions), numeric data (such as formant frequencies), categorical data (lexical items) or a matrix of differences obtained through the use of external software. The user can have a maximum of 20 projects on the server at any one time. Deleting a project is straightforward, and instantly frees up space for new maps and data.

3. INSPECTING THE DATA. Each project has its own entry screen giving access to 15 functions. The functions are subdivided into seven categories. The first four of these give an overview of the data and allow the user to check whether the data have been accurately represented by the algorithm. The first two functions are named *places* and *items*.

The *places* rubric allows inspection of the geographic locations that are associated with the linguistic data. In the example presented here, the locations have been given the names of languages and dialects that are to be compared. The polygon map submitted by the user is split into regions by Gabmap. It may happen that locations are unevenly distributed over the map. In this case a button on the data upload screen forces Gabmap to shift locations sufficiently to allow a clearer picture. In figure 2 below, this function was activated since some of the dialects of Kaska and Dena'ina lie at locations that are very close to each other on the map. The map was drawn to include Alaska and western Canada only, since all the data are from locations in this geographic area. The map on the left shows the divisions created by the algorithm, while the map on the right identifies the associated languages through numbers.

[5] Finer detail in the lakes and Alaskan coastlines of this map are due to Christopher Cox's deft mouse control.

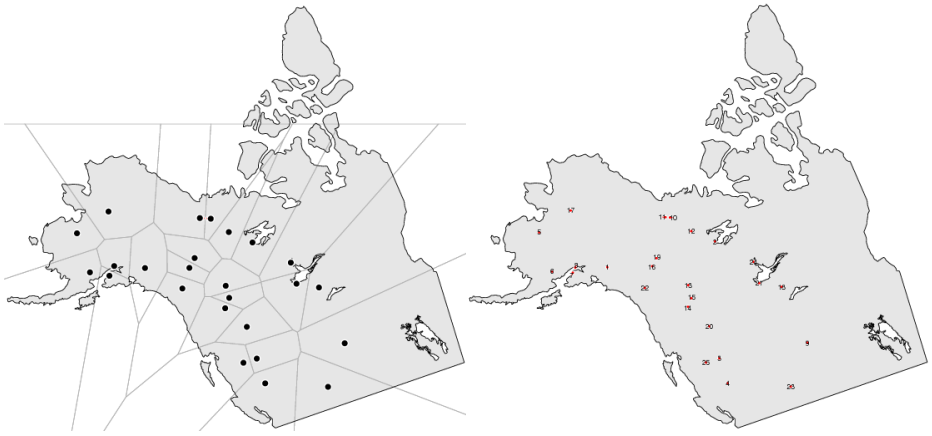


FIGURE 2. Athapaskan North America. Automatically generated areas on the left and languages on the right: 1 Ahtna, 2 Bearlake, 3 Central Carrier, 4 Chilcotin, 5 Deg Xinag, 6 Dena'ina (Inland), 7 Dena'ina (Outer Cook Inlet), 8 Dena'ina (Upper Cook Inlet), 9 Dene Sųłiné, 10 Gwich'in (Gwichya), 11 Gwich'in (Teetl'it), 12 Hare, 13 Kaska (Frances Lake), 14 Kaska (Good Hope Lake), 15 Kaska (Liard), 16 Kaska (Pelly), 17 Koyukon, 18 Mountain Slave, 19 Northern Tutchone, 20 Sekani, 21 South Slave (Hay River), 22 Southern Tutchone, 23 Tsuut'ina, 24 Dogrib, 25 Witsuwit'en.

The next function, *items*, lists all the concepts (or terms) that have been used in gathering the comparative data. In the case of the example data used here, this list comprised 52 body part terms. A color-coded map is displayed above the list, in which depth of hue is used to indicate the relative frequencies of the data. The number of forms used in the comparison of each concept cross-linguistically is also listed. For example, the term 'cheek' is represented by 21 forms in the sample, with data not available for four locations. These locations are identified by the color white in figure 3 below. The availability of items can also be investigated individually.

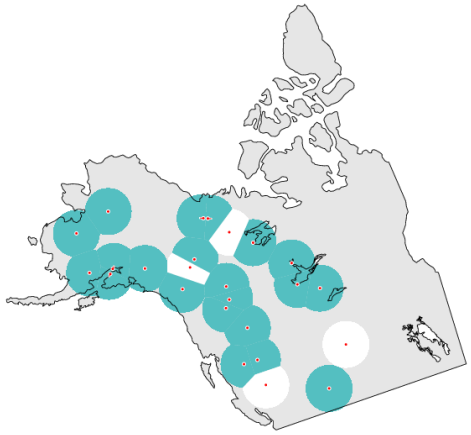


FIGURE 3. Availability of the item 'cheek', white areas indicate missing data

A further possibility for inspecting the structure of the sample comes through the function *data overview*. The function provides a list of the characters occurring in the sample along with their corresponding frequencies. Here, too, it is possible to inspect the geographic distributions of characters. Up to 200 strings containing the character in question are listed along with the map.

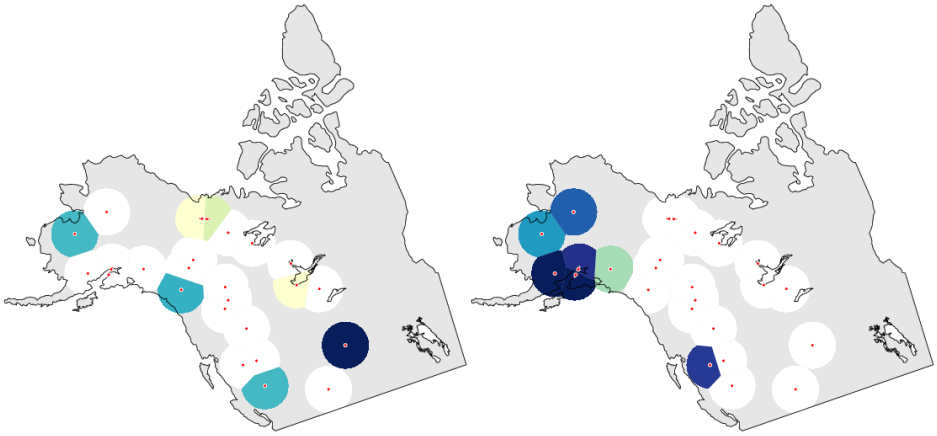


FIGURE 4. Distribution of characters /ð/ (left map) and /q/ (right map). White shading indicates absence of the character; darker hue indicates higher frequency

Beyond the geographical distributions themselves, these functions support the checking of data by making it easy to inspect the whole character set represented in the sample in the search for potential errors.

4. DISTRIBUTION MAPS. The development of Gabmap was chiefly motivated by research questions in dialectology. Consequently, one of the central aspects of Gabmap is the association of linguistic data with geographic locations. While the overall design is clearly aimed at quantitative evaluations of dialectological data, or dialectometry, all the functions described so far rely exclusively on mapping the data itself, without calculating measures of distance. As such, they can be useful tools for linguists interested in the distributions of forms and phonemes even when working with datasets that are perhaps too small to warrant quantitative evaluation. Alternatively, a particular researcher may be interested primarily in the distribution of individual lexical forms and isogloss boundaries. The function *distribution maps* operates without measuring aggregate distances, relying instead on the mapping of whole strings encoding individual concepts. A comparison can be made by selecting a concept (item), which brings up a display of all the phoneme strings that express this concept in the sampled languages. The researcher can then select a subset of strings to be displayed on the map, either by a point and click interface (keeping Control or Command pressed allows one to select multiple individual strings) or through a Regular Expression. Figure 5, for instance, shows the distribution of cognate terms for 'blood' (in dark blue) versus areas which have non-cognate expressions (in white). In this manner, the different realizations and lexicalizations of a concept can be represented in a series of

complementary maps. In the Athapaskan data below, one cognate set for 'blood' dominates, with five languages showing localized divergence.

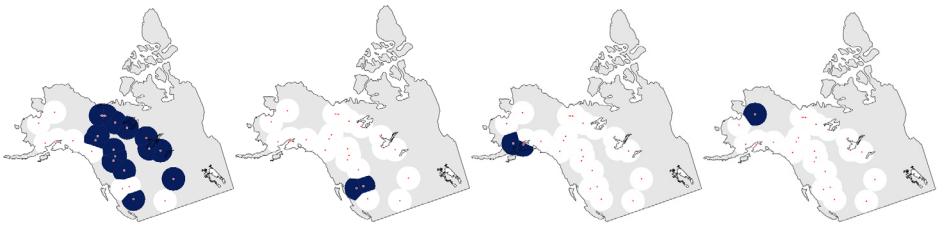


FIGURE 5. Distribution of cognate forms for the concept 'blood'

5. MEASURING DISTANCES. The remaining functions in Gabmap rely on the program's powerful capabilities for finding groups in the data. Specifically, the functions are concerned with how members of a set of languages or dialects relate to each other on the basis of a measure of distance. The raw data that is the input to Gabmap, in the form of phonetic strings for example, needs to be transformed into a series of distances measured between the objects, i.e. dialects or languages that are to be grouped. While several types of data are possible (as outlined in §2), the foremost distance measure that Gabmap relies on is the simple Levenshtein distance between strings representing the transcriptions of pronunciations. The Levenshtein distance between two words is informally defined as the minimum number of single-character edits required to change one string into another (Kruskal 1983: 18). Gabmap aligns the phoneme strings of the data so that vowels will be compared with vowels, and consonants with consonants. The distance between two strings is then established by comparing each character: if the characters are identical at an aligned location in each of the two strings, the distance will be measured as 0. If the two aligned characters are different the distance will be measured as 1. Should only a diacritical mark (e.g. /t/ vs. /t^h/) distinguish the two characters, the distance will be measured as 0.5. The distance between two strings is the sum of the character distances. In figure 6 below, two phoneme strings representing the concept 'thumb' are compared and found to have a Levenshtein distance of 3.5.

Ahtna	—	Kaska (Frances Lake)				
l	a		k^h	o	ts'	
l	aː	s	$ʧ^h$	o	ʔ	
0	0.5	1	1	0	1	3.5

FIGURE 6. Measuring the simple Levenshtein distance between two strings

It may be remarked here that Gabmap does not naturally treat digraphs and trigraphs as unitary symbols, such as in the case of the /ts'/ in figure 6. It was my choice to encode them in this manner based on phonological reasons specific to Athapaskan languages. Diacritical marks such as vowel length or aspiration are measured as distances of 0.5. So as not to

exaggerate the distance between longer words compared to shorter words, the "distance of each word pair is normalized by dividing it by the mean length of the word pair" (Nerbonne et. al 1999: x). The function *alignments* allows the user to inspect the arrangements and measurements of distance.[6]

String distance measurements produce a distance value between two languages based on one comparative concept. Of course, what is really desired is a measure based on all concepts in the comparison. This aggregated distance measure is obtained by taking the average value of all concept-based distances between each possible pairing of languages. All the measurements between individual language pairs are in a distance matrix. Table 1 shows a small part of such a matrix. The distance matrix created by Gabmap from the input data can be downloaded from the website.

	Ahtna	Bearlake	Central Carrier	Chilcotin
Ahtna	0	1138.41	1482.68	1716.69
Bearlake	1138.41	0	1199.73	1459.84
Central Carrier	1482.68	1199.73	0	269.677
Chilcotin	1716.69	1459.84	269.677	0

TABLE 1. Excerpt of distance matrix

6. DIFFERENCE MAPS. Gabmap implements several methods of data visualization that have been developed in dialectometry (see especially Goebl 2006, 2010 and Haimerl 2006). The first of these are beam maps (*Strahlenkarten*). The data going into this map are distance measures. Proximity is encoded in the map through depth of hue. Neighboring locations are connected by colored lines - or beams - showing the strength of association between the two locations.

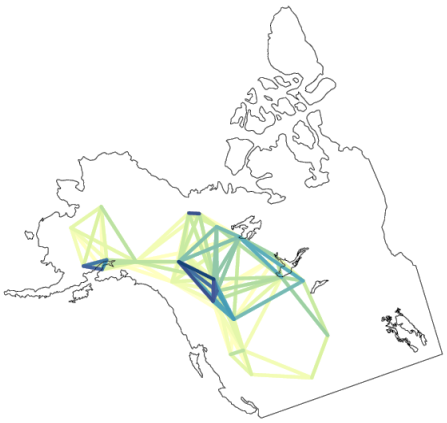


FIGURE 7. Beam map showing the strength of association between neighboring locations based on aggregate phonological distance.

[6] I must thank site administrator Çağrı Çöltekin who added the possibility of downloading sets of alignments as a text file at my request, providing highly useful additional functionality.

In Figure 7, very dark lines connect adjacent locations around the Cook Inlet in Alaska. These locations represent dialects of Dena'ina. In the north, a short blue line connects two dialects of Gwich'in. From a broader perspective the two Gwich'in locations are at the northern end of a dialect complex that spans the Canadian interior regions from the Slave and Great Bear lakes to the Rocky Mountains. The Kaska dialects, a more tightly knit group within this larger complex are connected by particularly dark lines. From a dialectometrists' perspective, the locations in figure 7 are likely still too sparse, but for the Athapaskanist they quite nicely delineate the Canadian interior dialect complex and its neighbors (Gwich'in in the north and Dene Sųłiné in the south).

The associations between locations can also be inspected at the level of individual locations. The function *reference point maps* (another functionality that goes back to the work of Haimerl 2006) gives a perspective on the proximity of the languages in the sample to one chosen language. For example, choosing the location for the Liard dialect of Kaska in figure 8, paints a map marking closely associated areas in dark blue. A black star marks the point of reference to which all the other languages are compared. Since Kaska forms part of the dialect complex delineated in the beam map in figure 7, many of the associated locations are dark in this map too.

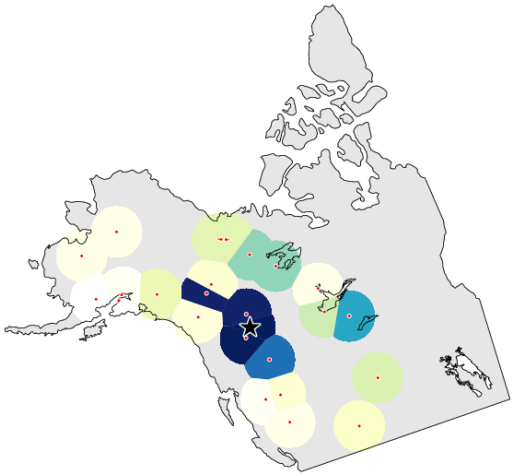


FIGURE 8. Depth of hue indicates relative distance to a chosen location, marked with a black star (here the Liard dialect of Kaska)

This function also produces a graph measuring the linguistic distance as a function of geographic distance. The plot on the left hand side of figure 9 shows the relationship between geographic distance and linguistic distance as seen from the perspective of the Liard dialect of Kaska. The four circles in the bottom left corner of the graph are represented as the darkest areas on the map in figure 8. The sharp increase in linguistic distance over the first few hundred kilometers is characteristic of what Nerbonne (2010:5) has called 'Séguy's Curve,' in honor of the French linguist who first observed it. This relationship between geographic and linguistic distance can be shown to hold generally true for the data in the sample, as evidence by the plot on the right in figure 9.

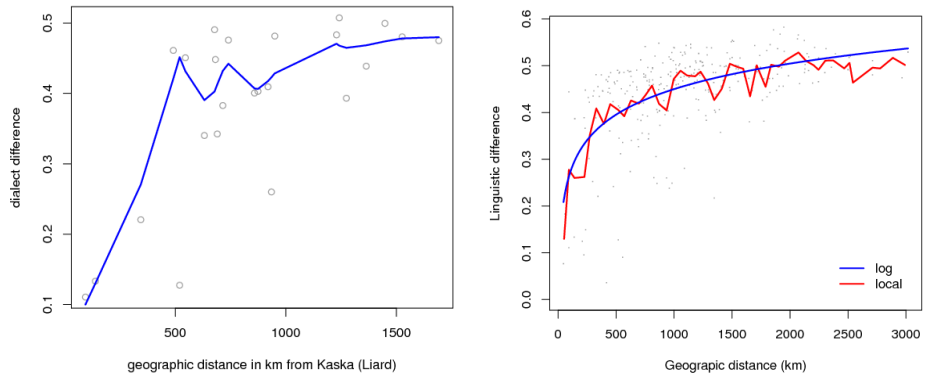


FIGURE 9. Measurements of linguistic distance (y-axis) over geographic distance (x-axis)

7. MULTIDIMENSIONAL SCALING. Multidimensional scaling (MDS) is a technique for representing distances between objects, as measured from a set of variables encoded in a distance matrix, to a set of relative positions in a "low dimensional multidimensional space" (Borg and Groenen 2005:3). MDS creates dimensions along which the objects to be compared are positioned, thereby reducing the full dimensionality of the original data. Reducing the data to two meaningful dimensions has the added benefit of making the resulting plot visually interpretable, as in figure 10.

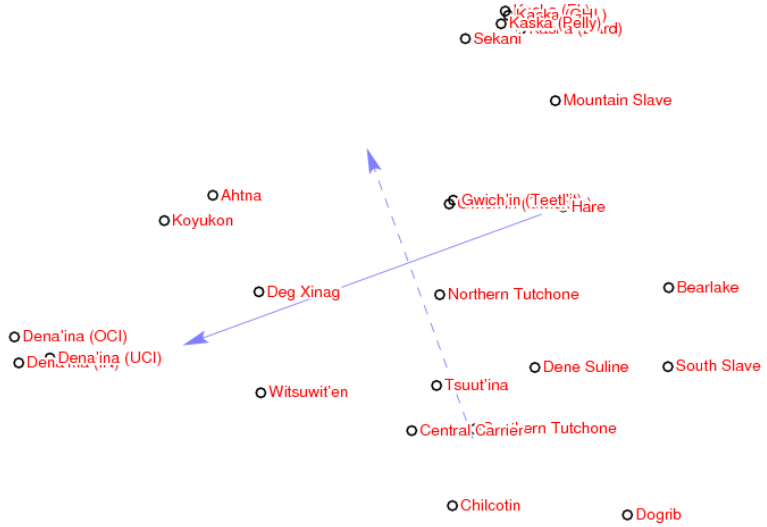


FIGURE 10. MDS plot of languages in the sample

The distortion of the data resulting from the reduction of the number of dimensions used in creating the plot can be measure with a *stress* value (that Gabmap also provides). The exact interpretation of this value is not immediately obvious, however. In fact, there is a

great deal to know about the functioning and uses of MDS, which are described in detail by Borg and Groenen (2005), and more briefly in Johnson (2008). It is important to note here that MDS provides a mathematical method for positioning the objects to be compared (the language or dialects) in a geometric space defined by two or more axes. The positioning for the languages in Figure 10 is not strictly representative of their geographic location, but of their relative position in terms of aggregate phonological distance. Nevertheless, it can be observed that the axis represented by the dashed arrow quite neatly divides Canadian from Alaskan languages (with the exception of Witsuwit'en). The Dena'ina and Kaska dialects cluster at the edges of the plot, while the right lower corner is somewhat sparsely populated by the languages of the Canadian interior, from Northern Tutchone in the Yukon to Tsuut'ina in the south.

8. CLUSTERING AND CLUSTER VALIDATION. MDS has a second very important function in Gabmap. Since it represents a robust assessment of relative distance among objects (Nerbonne et al. 2011:15), it is implemented as a cross-check on clustering. Nerbonne and colleagues warn of the danger of interpreting the highly appealing dendrograms derived through hierarchical agglomerative clustering too readily, since they are very sensitive even to small variations in the input data (Nerbonne et al. 2008, Kleiweg et al. 2004). Producing cluster dendrograms and plotting their results on maps is provided under the function entitled *cluster maps and dendrograms*. The languages represented in the distance matrix can be clustered according to one of four methods: weighted average, Ward's method, complete link, and group average. Each method results in slightly different cluster dendrograms. The dendrogram in figure 11 was constructed with the weighted average method. The numbers along the horizontal axis represent a measure of cophenetic distance. This measure can be used to estimate the distance between two languages from the dendrogram by examining the point at which the branches representing individual languages are joined to form clusters. For example, the cophenetic distance between Northern Tutchone and the two dialects of Gwich'in is at around 3.6 while the distance between the Pelly and Frances Lake (FL) dialects of Kaska is only 0.4. Bearlake and Mountain Slave are 'sisters' residing at the same level of the dendrogram, just like the two aforementioned dialects of Kaska. The latter, however, are much closer to each other as expressed visually in the length of the branches, and numerically in the cophenetic measure. Similarly, the distance between the cluster containing the Dena'ina dialects and the cluster containing Ahtna and Koyukon is 2.2. In sum, the length of the connecting branch in a cluster is an indication of the proximity of cluster members to each other, as well as of sub-clusters to other sub-clusters in a particular branching.

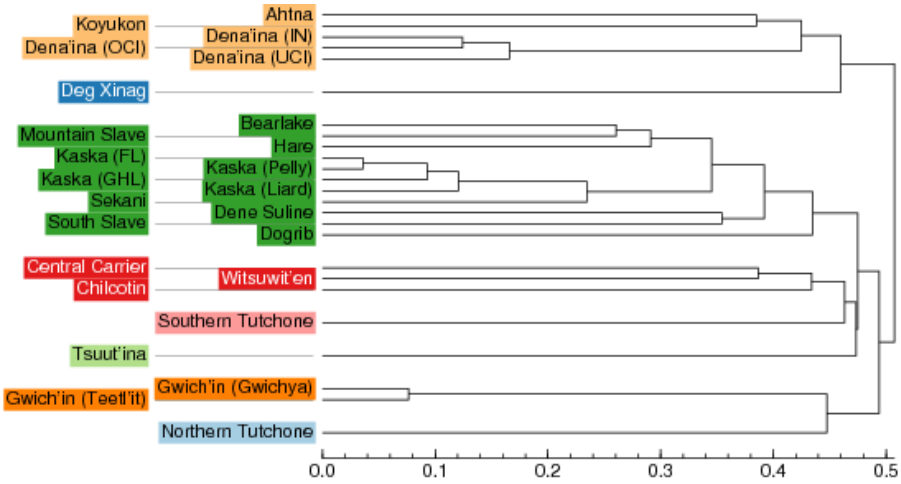


FIGURE 11. Cluster dendrograms of Northern Athapaskan languages

The dendrogram groups the languages in the sample into eight clusters. The clusters are color-coded so that their geographic distribution can be clearly identified on the *cluster map* (figure 12). The groupings that emerge from this clustering procedure are appealing from an Athapaskanist perspective because they correlate quite well with current understanding of sub-grouping in the family (Mithun 1999:345). The clustering does, however, group certain languages together that cut across distinctions made in the traditional classification.

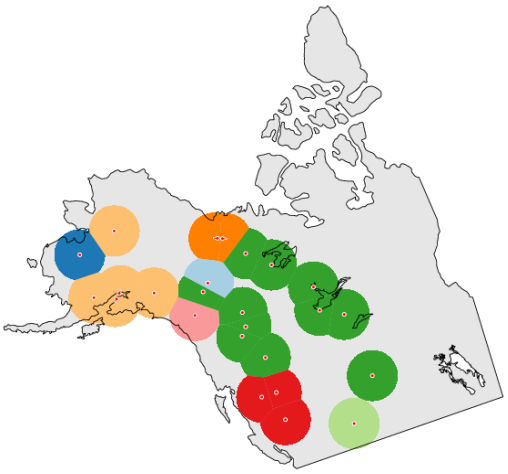


FIGURE 12. Cluster map of Northern Athopaskan languages

The languages called Southern Tutchone and Northern Tutchone are grouped together in the classification presented by Mithun, but in the dendrogram they belong to distinct clusters. Before jumping to further conclusions it is necessary to check the validity of indi-

vidual clusters. The most immediate means of doing this provided by Gabmap comes in the form of an MDS plot in which the clusters are represented as circles of matching color. In addition, both the cluster map and the MDS plot can be viewed in a black and white format, with the cluster being identified on both plots by being represented through matching symbols. The latter representational format is shown in figure 13 below.

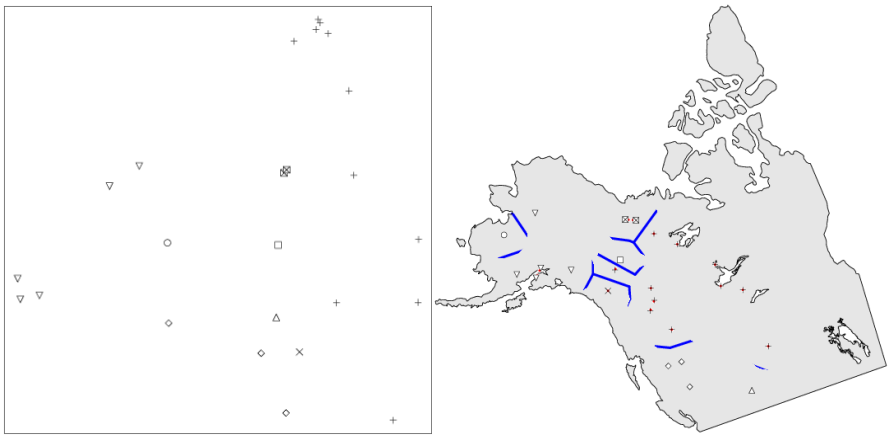


FIGURE 13. MDS plot of clusters from Figure 11

Gabmap provides a number of ways to inspect these clusters further in the MDS plot. The clusters can be manipulated at varying levels of resolution: certain clusters can be removed in order to show the division of multidimensional space among the remaining clusters. The number of clusters and the clustering method can also be adjusted. It is worth noting again that different methods can lead to differing structures among the clusterings. Adjusting the various settings and representations is easy and fast, inviting the user to explore their own data as well as the implications of various methodological choices.

Visual inspection of figure 13 already reveals that the clusters represented through downward pointing triangles (Alaskan languages) are more distant from all the other languages. The Alaskan languages around Cook Inlet form a stable cluster, with other Alaskan languages being distinct but close. This validates both the cluster and the sub-division of the cluster marked with the light orange color in the dendrogram. The crossed squares and the empty square (Gwich'in dialects and Northern Tutchone respectively) are still relatively close to each other, and to the long chain of crosses that here represent the green cluster of figure 11. The orange, light blue and green groupings of the dendrograms can also be interpreted as stable groupings, with the caveat that the green cluster (or crosses) represents a long dialect chain. The lower-center part of the MDS plot does not match any clusters particularly well, and it is not likely that the data provide any real indication for grouping here. Nevertheless, it may be remarked that the traditional association of Northern and Southern Tutchone with each other is contradicted by the data here, which see the former patterning with the interior Canadian dialect complex, and the latter being, perhaps, closer to the Athapaskan languages spoken in British Columbia.

A further means of constraining the instability of the clusterings is found under the rubric *fuzzy clustering*, a methodology developed for dialectometry by Peter Kleiweg and colleagues (Kleiweg et al. 2004). With this method random noise is added to the distance matrix at a level that can be specified by the user. This procedure is then repeated over 100 times, with the level of the randomly assigned noise varying within the specified limits (Nerbonne et al. 2008:5). The results can be plotted on a dendrogram in which the clusters are labeled with percentage values. The percentages indicate the frequency at which particular clusters re-emerge in repeated clusterings with the simulated noise in the data. The user can also specify the percentage level deemed acceptable for the representation of a cluster in the dendrogram, or the 're-emergence threshold.' Figure 14 shows a dendrogram constructed with this procedure displaying only clusters that have exceeded a re-emergence threshold of 80%.

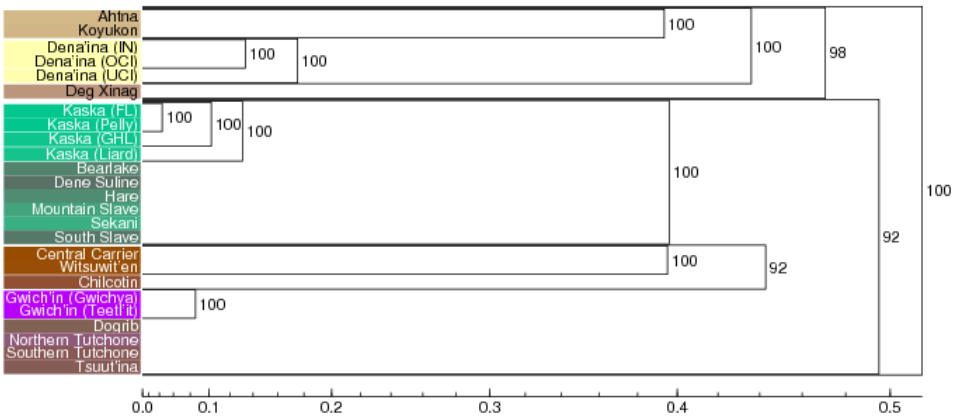


FIGURE 14. MDS plot of clusters from Figure 11

From this dendrogram, it can be observed that the overall two-way split between the Alaskan (light brown and yellow) languages and the Canadian languages is very stable. In the Alaskan group, the clusters indicating the structure of the internal relationships among the languages are valid: Ahtna and Koyukon are closer to each other and to the Dena'ina dialects than they are to Deg Xinag. The internal structure of the Canadian languages is more complicated. Clear divisions can be observed for the languages marked with green colors, and those in mauve and purple tones, but the internal structure of these clusters is not well differentiated. The exceptions are the dialects of Kaska and the languages of British Columbia (Central Carrier, Witsuwit'en, and Chilcotin), which emerge as stable, structured clusters. Thus there are four groups among the Canadian languages, but it is not obvious from the dendrogram how they are related to each other.

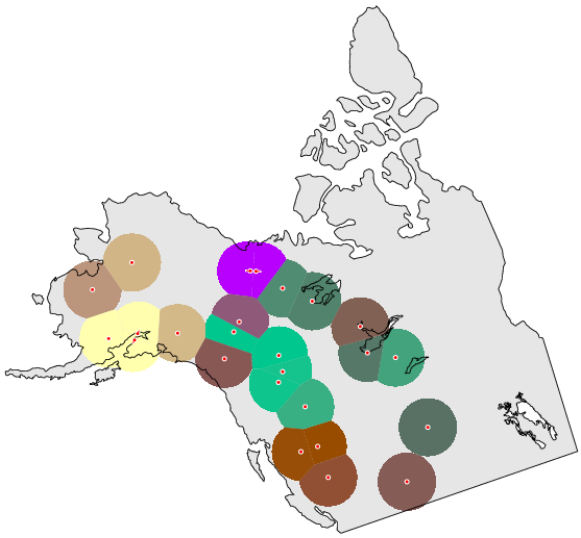


FIGURE 15. Fuzzy cluster map representing data in the dendrogram of Figure 14

The results of the dendrogram can be plotted on the map, as in figure 15. The colors indicate the groupings at the leaf level of the dendrogram. The geographic association of Alaskan grouping is confirmed once more, as is the grouping of languages spoken in British Columbia. The similarity in the hue of the areas in the south is deceiving, however, since there is really no indication that Tsuut'ina (represented by the brown circle straddling the bottom of the map) is in any way closely related to the British Columbia group. The automatic color assignment is, in this case, somewhat less than fortuitous.

8. CONCLUSION. Gabmap is excellent software that permits the mapping and comparison of linguistic data in a fast and generally painless manner. Since it is freely available and staffed by very responsive administrators it is to be hoped that usage of this software will grow. While extensive dialect data is not yet available for many endangered and lesser studied languages, the software can still be useful in the comparison of data, even if just for a small number of communities. The plotting of linguistic data on user-created maps, with its ample use of color, make it an appealing tool for data visualization. With Gabmap, complex linguistic data arrays can be presented in an easily understandable and visually stimulating manner. Gabmap can make computational tools for measuring linguistic relationships easily approachable and accessible, but it does not ultimately obviate the need for an understanding of the methods used, at a level of some sophistication.

Pros:

- Gabmap allows the association of linguistic and geographic data without the need for in-depth technical knowledge (such as might be required for some implementations in R, for example). The user needs to have understood only how to format the data and draw maps in Google Earth.
- The application of computationally complex procedures becomes straightforward, and intimate knowledge of mathematics is not required.
- The algorithms implemented in Gabmap carry out all computations automatically, leaving the user to inspect the results.
- Cluster validation can occur essentially through visual inspection.
- Visual representations of linguistic data can help communication between researchers with and without formal academic training.

Cons:

- While all the computation is carried out automatically and the user does not actually need to know the mathematics behind the applications, a general understanding of the methods involved is still highly recommended. For example, four different methods of clustering are made available, but choosing between them is not a trivial matter.
- The documentation provided does not address this issue. In general, the documentation and help functions are those aspects of Gabmap that still require the most work. The inquiring user is directed to work by Nerbonne, Kleiweg, Goebl and others for more in-depth information on the methods used and their implementation.

Primary function: To compare geographic locations in terms of their associated linguistic structure, especially with regard to pronunciations.
Platforms: Any. Gabmap is a web-based implementation.
Reviewed version: Online in March 2014.
Documentation: A paper giving an overview of the software and its development is available on the webpage, along with a detailed tutorial. Help pages are distributed throughout the application, but are incomplete in many case and awaiting further development.

References

Borg, Ingwer & Patrick J. F. Groenen. 2005. *Modern multidimensional scaling*. Berlin: Springer.

Goebl, Hans. 2005. Dialektometrie. In Köhler, Reinhard, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik / Quantitative linguistics: Ein internationales Handbuch / An international handbook*, 498-531. Berlin: De Gruyter.

Goebl, Hans. 2006. Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21(4). 411-435.

Goebl, Hans. 2010. Dialectologia: Theoretical prerequisites, practical problems, and practical applications. *Dialectologia* 1 (Special Issue I). 63-77.

Google Developers. 2013. *Keyhole Markup Language*. https://developers.google.com/kml/documentation/. (4th March, 2014.)

Haimerl, Edgar. 2006. Database design and technical solutions for the management, calculation, and visualization of dialect mass data. *Literary and Linguistic Computing* 21(4). 437-444.

Johnson, Keith. 2008. *Quantitative methods in linguistics*. Malden, MA: Blackwell Publishing.

Kleiweg, Peter, John Nerbonne & Leonie Bosveld. 2004. Geographic projection of cluster composites. In Blackwell, Alan, Kim Marriott & Atsushi Shimojima (eds.), *Diagrammatic representation and inference*, 392-394. Berlin: Springer.

Krauss, Michael E. & Victor Golla. 1981. Northern Athapaskan languages. In William C. Sturtevant (ed.), *Handbook of North American Indians: Subarctic*, 67-85. Washington: Smithsonian Institution.

Kruskal, Joseph B. 1983. An overview of sequence comparison. In David Sankoff & Joseph B. Kruskal (eds.), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*, 1-40. London: Addison-Wesley.

Mithun, Marianne. 1999. *The languages of native North America*. Cambridge: Cambridge University Press.

Nerbonne, John. 2010. Measuring the Diffusion of Linguistic Change. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365. 3821-3828.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg & Therese Leinonen. 2011. *Gabmap — a web application for dialectology*. http://www.gabmap.nl. (1st March 2014.)

Nerbonne, John, Peter Kleiweg, Franz Manni & Wilbert Heeringa. 2008. Projecting dialect distances to geography: Bootstrap clustering vs. noisy clustering. In Preisach, Christine, Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker (eds.), *Data analysis, machine learning, and applications. Proceedings of the 31st annual meeting of the German Classification Society*. Berlin: Springer.

Nerbonne, John, Wilbert Heeringa & Peter Kleiweg. 1999. Edit distance and dialect proximity. In Sankoff, David & Joseph B. Kruskal (eds.), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*, 2nd edn, v-xv. Stanford, CA: CSLI Publications.

Open Geospatial Consortium. 2014. *Standards: KML*. http://www.opengeospatial.org/standards/kml/. (4th March, 2014.)

Conor Snoek
snoek@ualberta.ca